

# Weekly Report

2013.5.6-2013.5.12

黄芯芯

## 本周工作：

1. 组会报告。
2. 看了一些关于个性化推荐的文章，一些总结和摘录如下：

个性化推荐系统三个核心模块：

### ● 用户特征提取模块

用户特征包括：年龄、性别、搜索过的关键词、浏览过的网页，还有一些评分、评论、收藏等网络行为等。

除了年龄性别之外，其他特征往往维度较高，因此很多推荐系统都会使用降维的方法以降低特征的维度。

第一种降维思路是根据兴趣将用户进行聚类。例如根据用户的历史兴趣将用户分成不同的类别，如历史爱好者、科技爱好者等。

聚类算法两种：硬聚类（每个对象只能属于一个类别，如 k-means 和层次聚类）和软聚类（每个对象可以属于多个类别，如 topic model）。

推荐系统中另外一个著名的模型：隐因子模型（Latent Factor Model,）

### ● 相关物品检索模块

获得用户兴趣特征后，根据特征找到和用户相关的物品。

利用信息检索中的成熟技术：倒排检索

计算“特征-物品”索引

推荐系统中最著名的所以是“物品-物品”索引，如何计算物品之间的相关度一直是推荐系统研究的重要问题之一。

两种计算物品相关度的算法：基于内容的物品相关度算法（利用自然语言处理和文本挖掘等）和基于用户行为的物品相关度算法（collaborative filtering 协同过滤，主要思想：如果喜欢物品 i 的用户也喜欢物品 j，那么 i 和 j 就有比较大的相关度，数据挖掘中的关联规则）（数据挖掘方法的缺陷在哪里？可视分析如果改进或辅助？）

### ● 推荐结果排序模块

需要一个统一的标准来计算用户对物品的兴趣，对物品进行排序。统一的指标最好能够满足两个条件：反映使用推荐系统的网站的商业目标；便于现有的成熟的机器学习法直接优化。使用两类分类器（逻辑回归 Logistic Regression，支持向量机 SVM，决策树促进树，神经网络等）

3. 关于淘宝项目，指导了一下建霞的毕业设计并着手研究陈畅那边的电信数据。

## 下周计划：

1. 研究电信数据。